

Project Final Report: Analyzing Real Estate Market Dynamics in Austin, TX

Sameen Siddiqua

1. Background:

The dataset offers detailed insights into residential properties across Austin, TX, encompassing various property features, location details, amenities, and pricing information. It comprises a mix of quantitative and categorical variables, enabling a comprehensive exploration of the local housing market. This dataset is a valuable asset for analyzing housing trends, property valuation, and market dynamics in Austin. It supports diverse analyses like hypothesis testing, regression modeling, and market segmentation based on property attributes. Furthermore, it helps uncover key factors influencing property prices and buyer preferences in the Austin real estate landscape.

1.2 Questions:

1. What are the key factors that significantly impact property prices in the housing market, and how do these factors interact with specific property characteristics to shape market dynamics?

1.3 Project Significance:

This project holds significant value in understanding real estate market dynamics and guiding data-driven decision-making. By applying statistical models such as independent t-tests, chi-square tests, logistic regression, and linear regression to analyze factors such as property tax rates, home types, garage presence, scenic views, and city locations, this research contributes to a deeper understanding of what influences property prices and desirability.

Key Significance Points:

- **Real Estate Market Insights:** Gain insights into factors influencing property prices, benefiting homeowners, real estate professionals, and policymakers.
- **Data-Driven Decision Making:** Demonstrates the power of statistical analysis in predicting real estate trends, guiding investment decisions, and informing development strategies.
- **Identifying Desirable Property Characteristics:** Highlights which property features contribute most to attractiveness and marketability through logistic regression analysis.
- **City-Level Comparisons:** Reveals geographic variations in property markets using chi-square tests, aiding urban planning and development strategies.
- **Practical Applications:** Provides actionable insights for homebuyers, sellers, investors, and policymakers based on t-tests and regression analysis results.

- **Research Contribution:** Contributes to the body of knowledge on real estate economics, inspiring further research in housing market dynamics using advanced statistical methods.

This project's findings offer practical implications for stakeholders and add to the broader research on real estate economics by leveraging various statistical models to analyze and interpret real estate data effectively.

2.Data Analysis:

2.1 About the Data Set

The dataset contains 47 variables that provide detailed information about residential properties in the Austin, TX area. Here is a summary of the key variables and their potential meanings:

1. Quantitative Variables:

- **latitude:** Latitude coordinates of the property location (continuous).
- **longitude:** Longitude coordinates of the property location (continuous).
- **propertyTaxRate:** Property tax rate applicable to the property (continuous).
- **garageSpaces:** Number of garage spaces associated with the property (discrete).
- **parkingSpaces:** Number of parking spaces available with the property (discrete).
- **yearBuilt:** Year the property was built (discrete)(year).
- **latestPrice:** Latest price of the property (continuous).
- **numPriceChanges:** Number of price changes for the property (discrete).
- **latest_saleyear:** Year of the latest property sale (discrete).
- **numOfPhotos:** Number of photos available for the property (discrete).
- **lotSizeSqFt:** Size of the property lot in square feet (continuous).
- **livingAreaSqFt:** Living area of the property in square feet (continuous).
- **numOfPrimarySchools:** Number of primary schools nearby (discrete).
- **numOfElementarySchools:** Number of elementary schools nearby (discrete).
- **numOfMiddleSchools:** Number of middle schools nearby (discrete).
- **numOfHighSchools:** Number of high schools nearby (discrete).
- **avgSchoolDistance:** Average distance to schools from the property (continuous).
- **avgSchoolRating:** Average rating of nearby schools (continuous).
- **avgSchoolSize:** Average size of nearby schools (continuous).
- **MedianStudentsPerTeacher:** Median number of students per teacher in nearby schools (continuous).
- **numOfBathrooms:** Number of bathrooms in the property (discrete).
- **numOfBedrooms:** Number of bedrooms in the property (discrete).
- **numOfStories:** Number of stories (floors) in the property (discrete).

2. Qualitative Variables:

- **city:** City where the property is located (nominal).
- **streetAddress:** Specific Street address of the property (nominal).
- **zipcode:** ZIP code of the property location (nominal).

	<ul style="list-style-type: none"> • description: Detailed description of the property (nominal). • homeType: Type of home (nominal). • latest_saledate: Date of the latest property sale (nominal). • latestPriceSource: Source of the latest price information (nominal). • homeImage: Filename or link to the image of the property (nominal).
3.	Boolean Variables:
	<ul style="list-style-type: none"> • hasAssociation: Indicates whether the property is part of a homeowners' association (TRUE/FALSE). • hasCooling: Indicates whether the property has a cooling system (TRUE/FALSE). • hasGarage: Indicates whether the property has a garage (TRUE/FALSE). • hasHeating: Indicates whether the property has a heating system (TRUE/FALSE). • hasSpa: Indicates whether the property has a spa (TRUE/FALSE). • hasView: Indicates whether the property has a view (TRUE/FALSE).
4.	Ordinal Variables:
	<ul style="list-style-type: none"> • None

These variables provide comprehensive details about the properties in the dataset, making it suitable for various types of analysis and modeling related to real estate and property valuation in the Austin area.

2.2 Data Preprocessing:

Data Collection: The dataset used in this project was obtained from Kaggle, a reputable source known for hosting high-quality datasets, ensuring the reliability and accuracy of our analysis.

Data Cleaning: The dataset was already clean, well-structured, and free of missing values.

Attribute Selection: We performed attribute selection to identify relevant features that contribute significantly to our analysis, streamlining the dataset to focus on key variables of interest.

Encoding Categorical Variables: Categorical variables, including boolean features, were encoded for modeling purposes. Boolean variables were converted into numeric format (1 for True, 0 for False) using Excel functions like **IF**, facilitating their integration into our analysis.

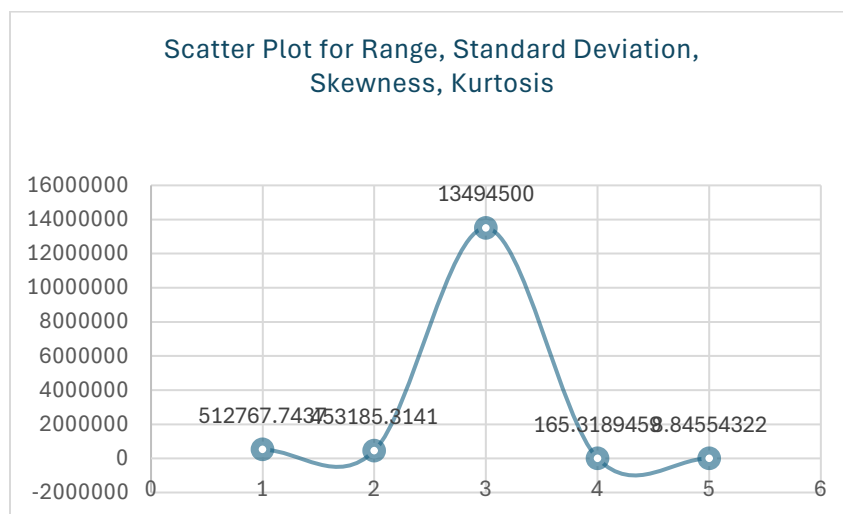
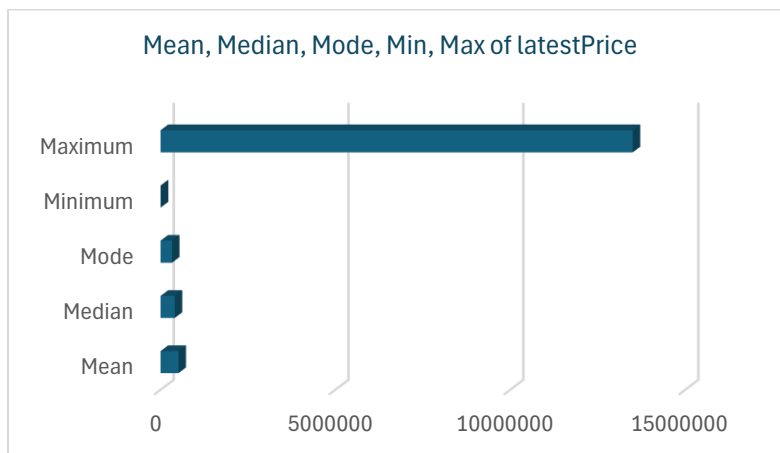
2.3 Understanding the Data Set:

A general understanding of the data set can be obtained by Descriptive statistics

latestPrice	
Mean	512767.7437
Standard Error	3679.329847
Median	405000
Mode	325000
Standard Deviation	453185.3141
Sample Variance	2.05377E+11
Kurtosis	165.3189459
Skewness	8.84554322
Range	13494500
Minimum	5500
Maximum	13500000
Sum	7779199440
Count	15171
Largest(1)	13500000
Smallest(1)	5500
Confidence Level(95%)	7211.929199

- **Mean:** The average **latestPrice** is approximately \$512,767.74, reflecting the central tendency of property prices in the dataset.
- **Median:** The median **latestPrice** of \$405,000 indicates that half of the property prices are below this value, providing a reliable measure of central tendency less influenced by extreme values.
- **Mode:** The most frequent **latestPrice** value (mode) is \$325,000, indicating the price point that appears most frequently in the dataset.
- **Standard Deviation:** The standard deviation of approximately \$453,185.31 shows substantial variability in **latestPrice** values around the mean, highlighting the range of property prices.
- **Range:** The **latestPrice** range spans from \$5,500 to \$13,500,000, demonstrating the diversity of property prices in the dataset.
- **Skewness:** The skewness of approximately 8.85 indicates a significant positive skew in the **latest Price** distribution, meaning the data is highly skewed towards higher prices. This skewness is supported by the fact that the mean **latestPrice** (around \$512,767.74) is greater than the median **latestPrice** (\$405,000). This discrepancy suggests that a subset of properties with very high prices (outliers) is pulling the average upwards, emphasizing the prevalence of higher-priced properties in the dataset.
- **Kurtosis:** The kurtosis value of 165.32 indicates extreme peakedness or outliers in the distribution of **latestPrice** values, suggesting a heavy-tailed distribution with many extreme values.
- **Confidence Level (95%):** The confidence interval at a 95% confidence level is approximately \$7,211.93, providing an estimate of the precision of the mean **latestPrice** value.

•

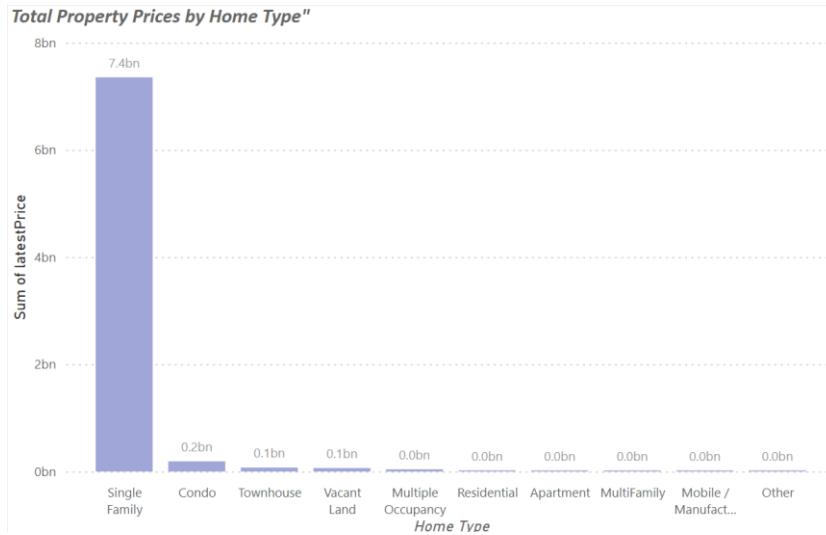


Descriptive Summary of the Dataset:

lotSizeSqFt		livingAreaSqFt	
Mean	119083.9681	Mean	2208.323314
Standard Error	99477.61434	Standard Error	11.19241093
Median	8276	Median	1975
Mode	11325.6	Mode	2068
Standard Deviation	12252718.78	Standard Deviation	1378.576119
Sample Variance	1.50129E+14	Sample Variance	1900472.117
Kurtosis	15142.40078	Kurtosis	2413.503728
Skewness	122.9985459	Skewness	32.40182912
Range	1508482700	Range	108992
Minimum	100	Minimum	300
Maximum	1508482800	Maximum	109292
Sum	1806622880	Sum	33502473
Count	15171	Count	15171
Largest(1)	1508482800	Largest(1)	109292
Smallest(1)	100	Smallest(1)	300
Confidence Level(95%)	194988.0933	Confidence Level(95%)	21.9384721

numOfBathrooms		numOfBedrooms	
Mean	2.683010349	Mean	3.440379672
Standard Error	0.00856085	Standard Error	0.006900524
Median	3	Median	3
Mode	2	Mode	3
Standard Deviation	1.054445135	Standard Deviation	0.849941772
Sample Variance	1.111854543	Sample Variance	0.722401016
Kurtosis	21.83036541	Kurtosis	11.50210777
Skewness	1.799876111	Skewness	0.801147731
Range	27	Range	20
Minimum	0	Minimum	0
Maximum	27	Maximum	20
Sum	40703.95	Sum	52194
Count	15171	Count	15171
Largest(1)	27	Largest(1)	20
Smallest(1)	0	Smallest(1)	0
Confidence Level(95%)	0.016780296	Confidence Level(95%)	0.013525857

3. VISUALS:



Insights: The total property price for Single Family homes is significantly higher than any other type of home, reaching up to approximately 8 billion. This suggests that single-family homes are the most prevalent type of property on the market. All other home types such as Condo, Townhouse, Vacant Land, Multiple Occupancy, Apartment, Multifamily, Mobile/Manufactured, and Other have much lower total property prices in comparison. This could indicate that these types of properties are less common or have lower individual prices.

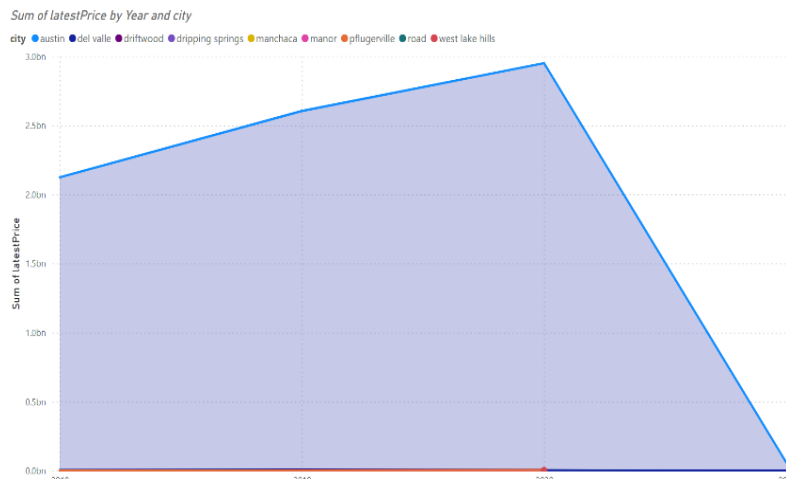


Insights:

The plot shows a collection of points that appear to be positively correlated; as the latest Price increases, the living area SqFt also tends to increase. This suggests that properties with higher prices tend to have larger living areas.

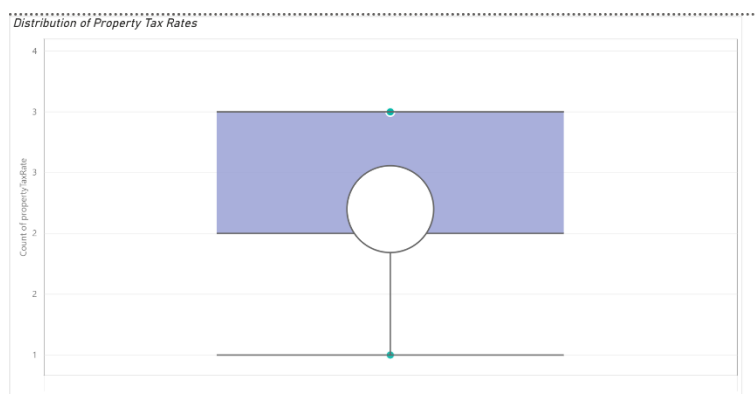
There are more data points clustered at the lower end of both axes, indicating that there are more properties with lower prices and smaller square footage.

A few outliers can be observed at higher price points with relatively large square footage. These could represent luxury properties or properties in high-demand areas.



Insights:

1. **Austin's Growth and Decline:** Austin had a significant increase in the price from 2018 to 2019, but then experienced a sharp decline in 2021. This could indicate a boom and subsequent cooling in the real estate market or other economic factors affecting property prices in Austin.
2. **Stability in Manchaca:** Manchaca had the same price in 2018 and 2019, indicating stability in property prices during this period.
3. **Increase in Manor:** Manor had a significant increase in price from 2018 to 2019, which could suggest a growing demand for properties in this area.
4. **Decline in Pflugerville Road:** Pflugerville Road had a decline in the price from 2018 to 2021, which could be due to a decrease in property prices or a reduction in the number of properties sold.



Insights:

The large white circle in the center of the graph likely represents data that falls within the most common range of property tax rates. This could be interpreted as the mode of the dataset, which is the most frequently occurring value(s). The mode is 2.20. The shaded rectangular area extending horizontally across the graph suggests the spread or distribution of property tax rates across different counties. The width of the shaded area at any given point could indicate the frequency or probability density of the property tax rates at that point.

4. METHODS FOR DATA ANALYSIS:

4.1 Hypothesis Testing

a. Hypothesis Test: Association between Home Type and Garage Presence

Background

A hypothesis test is conducted to explore whether there exists a statistically significant association between the type of home (homeType) and the presence of a garage (hasGarage) in a real estate dataset. The purpose of this analysis is to investigate whether certain types of homes are more likely to have garages compared to others.

Hypotheses

- Null Hypothesis (H0): The type of home (homeType) and the presence of a garage (hasGarage) are independent of each other.
- Alternative Hypothesis (H1): The type of home (homeType) and the presence of a garage (hasGarage) are not independent; there is an association between them.

Test Method: Chi-Square Test for Independence

To assess these hypotheses, the Chi-Square Test for Independence is utilized. This test is particularly suited for analyzing categorical data and determining if there is an association between two categorical variables.

Chi-Square Test Formula

The Chi-Square statistic is computed using the following formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \chi^2 = \sum E_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}^2}$$

Where:

- O_{ij} = Observed frequency count in a specific cell of the contingency table.
- E_{ij} = Expected frequency count assuming independence of the variables.
- i and j represent categories of homeType and hasGarage.

b. Hypothesis Test: Comparison of Property Prices Based on Tax Rates

Background

This hypothesis test aims to determine if there is a significant difference in the mean property tax rates (**propertyTaxRate**) between **Single Family** homes and **Town** homes in a given real estate dataset.

Hypotheses:

- **Null Hypothesis (H0):** There is no significant difference in the mean property tax rates between **Single Family** homes and **Town** homes.
- **Alternative Hypothesis (H1):** There is a significant difference in the mean property tax rates between **Single Family** homes and **Town** homes.

Test Method: Independent Samples t-Test

The Independent Samples t-Test is used to compare the means of **propertyTaxRate** between **Single Family** homes and **Town** homes. This test assesses whether there is a statistically significant difference in property tax rates between these two types of homes in the dataset.

4.2 Regression:

Regression is a statistical method used to understand and quantify the relationship between one dependent variable (often denoted as Y) and one or more independent variables (often denoted as X_1, X_2, \dots, X_p). The primary goal of regression analysis is to predict the value of the dependent variable based on the values of the independent variables. I have utilized both Multi Linear Regression and Logistic Regression to analyze different aspects of a real estate dataset.

a. Multi Linear Regression: Factors Influencing Home Prices

Dependent Variable: latestPrice (Price of homes)

Independent Variables:

- lotSizeSqFt (Lot size in square feet)
- livingAreaSqFt (Living area size in square feet)
- numOfBathrooms (Number of bathrooms)
- numOfBedrooms (Number of bedrooms)
- numOfStories (Number of stories)
- hasCooling (Binary indicator for cooling system presence)
- hasGarage (Binary indicator for garage presence)
- hasHeating (Binary indicator for heating system presence)
- numOfAppliances (Number of appliances)

Objective: To identify factors influencing home prices using a Multi Linear Regression model

Multi Linear Regression Equation:

The Multi Linear Regression equation can be represented as:

$$\text{latestPrice} = \beta_0 + \beta_1 \times \text{lotSizeSqFt} + \beta_2 \times \text{livingAreaSqFt} + \beta_3 \times \text{numOfBathrooms} + \beta_4 \times \text{numOfBedrooms} + \beta_5 \times \text{numOfStories} + \beta_6 \times \text{hasCooling} + \beta_7 \times \text{hasGarage} + \beta_8 \times \text{hasHeating} + \beta_9 \times \text{numOfAppliances}$$

$$\text{testPrice} = \beta_0 + \beta_1 \times \text{lotSizeSqFt} + \beta_2 \times \text{livingAreaSqFt} + \beta_3 \times \text{numOfBathrooms} + \beta_4 \times \text{numOfBedrooms} + \beta_5 \times \text{numOfStories} + \beta_6 \times \text{hasCooling} + \beta_7 \times \text{hasGarage} + \beta_8 \times \text{hasHeating} + \beta_9 \times \text{numOfAppliances}$$
 Where:

- β_0 is the intercept (constant term),
- $\beta_1, \beta_2, \dots, \beta_9$ are the coefficients of the independent variables.

4.3 Correlation Analysis

Correlation is a statistical measure that quantifies the strength and direction of the relationship between two variables, ranging from -1 to +1, where a coefficient closer to +1 indicates a strong positive relationship (both variables tend to increase or decrease together), a coefficient closer to -1 indicates a strong negative relationship (one variable increases as the other decreases), and a coefficient near 0 suggests little to no linear relationship between the variables.

5 . Results:

5.1 Hypothesis Testing:

a. Chi Square Test

Null Hypothesis (H0): The type of home (homeType) and the presence of a garage (hasGarage) are independent.

Alternative Hypothesis (H1): The type of home (homeType) and the presence of a garage (hasGarage) are not independent (i.e., there is an association between them).

Observed Values:


	Has garage(true)	Has garge(false)	Total
Apartment	18	19	37
Condo	261	209	470
Mobile/Manufactured	8	9	17
MultiFamily	5	5	10
Multi Occupancy	47	0	47
Other	5	1	6
Residential	31	6	37
Single Family	7836	6405	14241
Townhouse	106	68	174
Vacant Land	29	54	83
Total	8346	6776	15122

Expected Values:

	Has garage(true)	Has garge(false)	Total
Apartment	20.42071155	16.57928845	37
Condo	259.3982277	210.6017723	470
Mobile/Manufactured	9.382489089	7.617510911	17
MultiFamily	5.519111229	4.480888771	10
Multi Occupancy	25.93982277	21.06017723	47
Other	3.311466737	2.688533263	6
Residential	20.42071155	16.57928845	37
Single Family	7859.766301	7859.766301	15719.53
Townhouse	96.03253538	77.96746462	174
Vacant Land	45.8086232	37.1913768	83
Total	8346	8254.532602	16600.53

$(O-E)^2/E$

	Has garage(true)	Has garge(false)
Apartment	0.286955936	0.35344366
Condo	0.009890871	0.012182587
Mobile/Manufactured	0.20370672	0.250905591
MultiFamily	0.048826062	0.060139066
Multi Occupancy	17.09846164	21.06017723
Other	0.86099146	1.060483282
Residential	5.480775924	6.750672353
Single Family	0.071864357	269.2630937
Townhouse	1.03454887	1.274253965
Vacant Land	6.167611993	7.59664842

Df		9 (n-1)*(m-1)
Chi Square		338.9456337
P value		1.39349E-67
Alpha(0.05)		

Interpretation:

1. Chi-Square Value:

- The calculated Chi-Square statistic is 338.9456, which is a measure of the difference between the observed and expected frequencies under the null

hypothesis. A higher Chi-Square value indicates a stronger deviation from independence.

2. P-Value:

- The extremely small P-value (1.39349×10^{-67}) suggests strong evidence against the null hypothesis. This low P-value indicates that it is highly unlikely to observe such an extreme association between home type and garage presence if they were truly independent.

conclusion: Based on the statistical analysis:

- With a very low P-value ($p < 0.05$), we reject the null hypothesis.
- There is sufficient evidence to conclude that there is a statistically significant association between the type of home (homeType) and the presence of a garage (hasGarage).
- Therefore, we accept the alternative hypothesis (H1) that the type of home and garage presence are not independent; they are associated with each other.

b. Independent T Test

Null Hypothesis (H0):

The null hypothesis states that there is no significant difference in the mean property tax rates (**propertyTaxRate**) between **Single Family** homes and **Town** homes.

Alternative Hypothesis (H1):

The alternative hypothesis contradicts the null hypothesis and suggests that there is a significant difference in the mean property tax rates (**propertyTaxRate**) between **Single Family** homes and **Town** homes.

T Test: Two-Sample Assuming Unequal Variances		
	propertyTaxRate(single family)	propertyTaxRate(townhouse)
Mean	1.994731409	1.982643678
Variance	0.002938427	0.000604531
Observations	14241	174
Hypothesized Me	0	
df	194	
t Stat	6.300606209	
P(T<=t) one-tail	9.73783E-10	
t Critical one-tail	1.652745917	
P(T<=t) two-tail	1.94757E-09	
t Critical two-tail	1.972267488	

Test Results:

- **t Statistic:** 6.300606209
 - The t-statistic measures the difference between the means of **propertyTaxRate** for **Single Family** and **Town** homes relative to the variability in the data.
- **Degrees of Freedom (df):** 194
 - The degrees of freedom represent the number of independent observations used to calculate the t-statistic.
- **P-Values:**
 - **One-Tail P-Value ($t \leq t$):** 9.73783×10^{-10} – 9.73783×10^{-10}
 - **Two-Tail P-Value ($t \leq t$):** 1.94757×10^{-9} – 1.94757×10^{-9}
 - The extremely small P-values indicate the probability of observing such extreme differences in mean property tax rates under the null hypothesis.

Interpretation:

1. **t Statistic:**
 - The calculated t-statistic of 6.300606209 indicates a significant difference between the mean property tax rates of **Single Family** and **Town** homes.
2. **P-Values:**
 - The small P-values ($p < 0.05$) suggest strong evidence against the null hypothesis, indicating that the observed difference in mean property tax rates is unlikely to occur by random chance.

Conclusion:

Based on the statistical analysis:

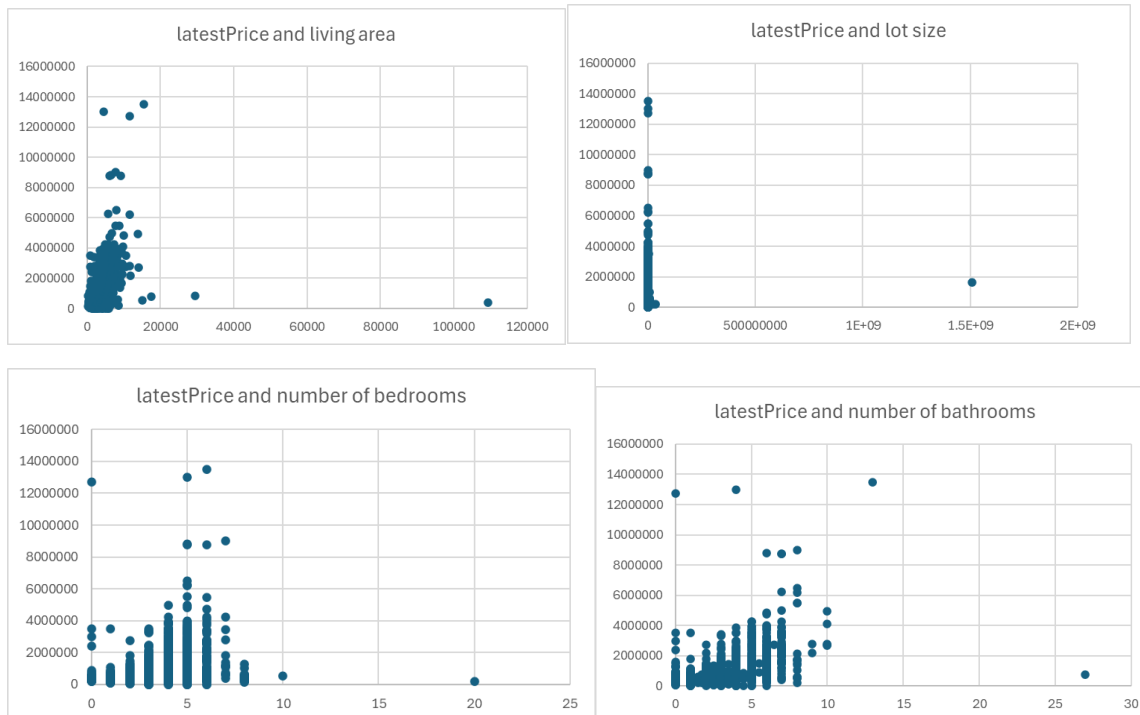
- The extremely small P-values ($p < 0.05$) lead us to reject the null hypothesis.
- There is strong evidence to suggest that there is a significant difference in mean property tax rates (**propertyTaxRate**) between **Single Family** homes and **Town** homes.

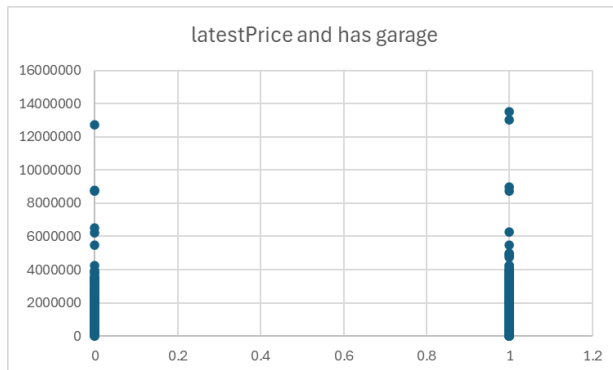
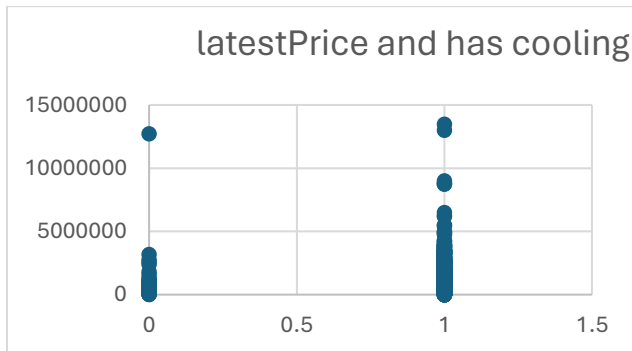
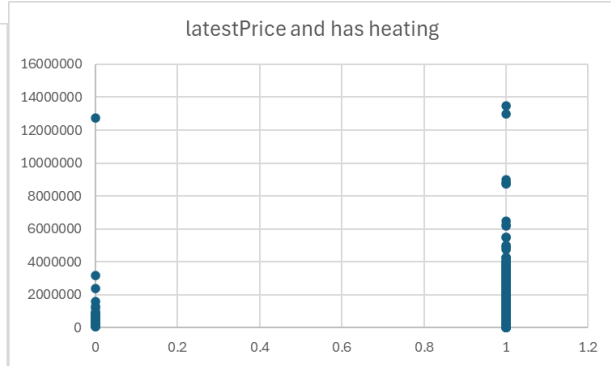
5.2 Regression:

a. Multi Linear Regression

SUMMARY OUTPUT									
Regression Statistics									
Multiple R		0.567746261							
R Square		0.322335817							
Adjusted R Square		0.321978258							
Standard Error		373162.1942							
Observations		15171							
ANOVA									
	df	SS	MS	F	Significance F				
Regression	8	1.0043E+15	1.25532E+14	901.4892227	0				
Residual	15162	2.1113E+15	1.3925E+11						
Total	15170	3.1156E+15							
		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept		206891.1566	32735.8672	6.320014539	2.68847E-10	142724.9	271057.4	142724.9	271057.3978
lotSizeSqFt		0.00056831	0.00024729	2.298141204	0.021567297	8.36E-05	0.001053	8.36E-05	0.00105303
livingAreaSqFt		95.56626492	2.75278552	34.71620451	3.5464E-254	90.17047	100.9621	90.17047	100.962056
numOfBathrooms		201465.1658	4385.65384	45.93731583	0	192868.8	210061.6	192868.8	210061.5754
numOfBedrooms		-46748.0675	4661.16551	-10.0292657	1.34279E-23	-55884.5	-37611.6	-55884.5	-37611.6219
numOfStories		-113370.8043	6954.60277	-16.3015499	3.06826E-59	-127003	-99738.9	-127003	-99738.9456
has cooling		-201971.8595	31217.2778	-6.46987418	1.01092E-10	-263161	-140782	-263161	-140782.236
has garage		34047.79423	6123.19647	5.560460851	2.73607E-08	22045.59	46050	22045.59	46049.99656
has heating		61610.13023	41897.6169	1.470492472	0.141449185	-20514.2	143734.5	-20514.2	143734.5039

Visuals:





Regression Model Summary:

- Multiple R: 0.5677
 - The multiple correlation coefficient (R) indicates a moderate positive relationship between the independent variables and the dependent variable (**latestPrice**).

- **R-Squared (R^2): 0.3223**
 - The coefficient of determination (R^2) suggests that approximately 32.2% of the variance in home prices (`latestPrice`) can be explained by the independent variables included in the model.
- **Adjusted R-Squared: 0.3220**
 - The adjusted R^2 adjusts the R^2 value to account for the number of predictors in the model, providing a more accurate reflection of the model's goodness-of-fit.
- **Standard Error: 373,162.1942**
 - The standard error of the estimate measures the average distance between the observed values and the predicted values by the model.

Analysis of Significant Predictors:

Significant Variables (p-value < 0.05)

1. **lotSizeSqFt** (p-value = 0.0216)
 - This variable is statistically significant at the 0.05 level, suggesting that the size of the lot (in square feet) has a significant impact on the housing price.
2. **livingAreaSqFt** (p-value = 3.55E-254)
 - Highly significant with an extremely low p-value, indicating that the living area (in square feet) strongly influences the housing price.
3. **numOfBathrooms** (p-value = 0)
 - This variable is highly significant with a p-value of 0, indicating that the number of bathrooms significantly affects the housing price.
4. **numOfBedrooms** (p-value = 1.34E-23)
 - Highly significant, suggesting that the number of bedrooms has a notable impact on housing price.
5. **numOfStories** (p-value = 3.07E-59)
 - Highly significant, indicating that the number of stories in the property is a significant predictor of housing price.
6. **has cooling** (p-value = 1.01E-10)
 - Statistically significant, implying that the presence of cooling systems affects the housing price.
7. **has garage** (p-value = 2.74E-08)
 - Significant at the 0.05 level, suggesting that having a garage influences the housing price.

Non-Significant Variable (p-value \geq 0.05)

1. **has heating** (p-value = 0.141)
 - This variable is not statistically significant at the 0.05 level, implying that the presence of heating systems may not significantly impact housing price based on this analysis.

Conclusion:

In summary, the variables lotSizeSqFt, livingAreaSqFt, numOfBathrooms, numOfBedrooms, numOfStories, has cooling, and has garage are all found to be statistically significant predictors of housing price in this regression model. Conversely, the presence of heating (has heating) does not appear to have a significant influence on housing price based on the given dataset and analysis.

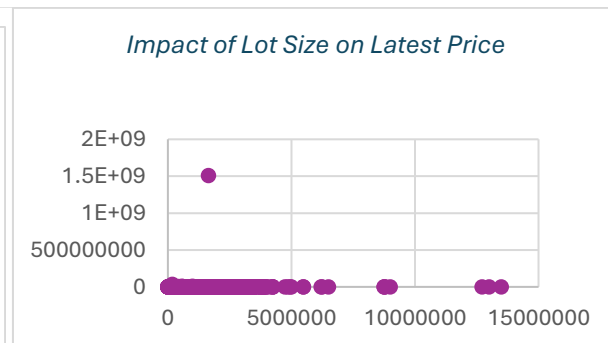
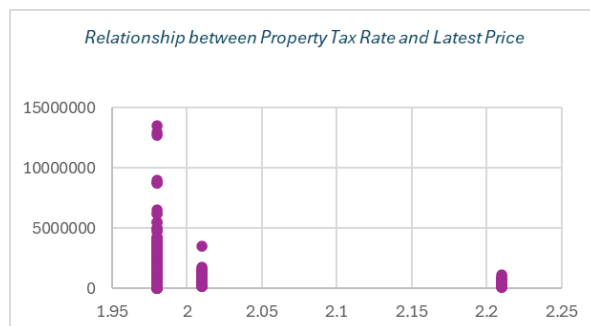
5.3 Correlation Analysis:

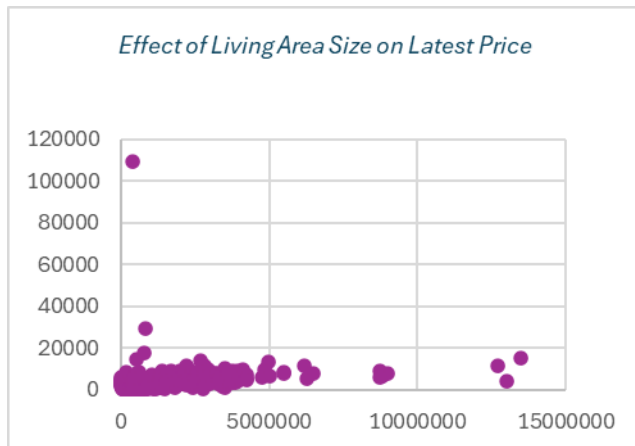
	propertyTaxRate	latest price	lotSizeSqFt	livingAreaSqFt
propertyTaxRate	1			
latest price	-0.062185085	1		
lotSizeSqFt	-0.001985405	0.020939531	1	
livingAreaSqFt	0.06245923	0.467031636	0.006953308	1

Interpretation:

1. **Property Tax Rate and Latest Price:** There is a weak negative correlation (-0.062) between Property Tax Rate and Latest Price, suggesting a slight tendency for higher property tax rates to be associated with lower property prices.
2. **Lot Size and Latest Price:** The correlation coefficient (0.0209) between Lot Size and Latest Price is very close to zero, indicating a negligible linear relationship between these variables.
3. **Living Area Size and Latest Price:** There is a moderate positive correlation (0.467) between Living Area Size and Latest Price, suggesting that larger living areas tend to be associated with higher property prices.

Visuals:





6. Future Work

1. **Exploring Additional Factors:** Investigate the impact of additional factors on property prices and desirability, such as proximity to amenities (parks, shopping centers), crime rates, and transportation accessibility.
2. **Machine Learning Approaches:** Implement machine learning algorithms (e.g., Neural Networks, Support Vector Machines) for more nuanced pattern recognition and predictive modeling.
3. **Market Segmentation:** Conduct market segmentation analysis to identify distinct buyer preferences and tailor marketing strategies accordingly.
4. **Data Collection and Integration:** Continuously update and integrate datasets to capture evolving market dynamics and ensure model robustness.

7.CONCLUSION

In conclusion, this project provides valuable insights into the real estate market dynamics in Austin, TX. Key findings highlight significant associations and predictors affecting property prices and desirability. The analysis demonstrates a strong relationship between home type and garage presence, along with notable differences in tax rates of single family homes and town homes. Factors such as living area size, bathrooms, bedrooms, stories, cooling system presence, and garage presence emerge as influential determinants of home prices. Moving forward, further exploration of additional factors and advanced modeling techniques will enhance predictive accuracy and market segmentation in real estate analysis.

8. Reference

1. <https://www.kaggle.com/datasets/ericpierce/austinhousingprices>
2. Business Analytics Sanjiv Jaggia